FACES, año 6, nº 7, enero/abril de 2000; págs. 83-106

Una aplicación de métodos multivariados: caracterización de los desocupados residentes en el Partido de General Pueyrredon

An application of multivariate statistical methods: unemployed residents in the General Pueyrredon District

Fernando Graña[*]
Anabel Marín[**]
Pablo Angelelli[**]

RESUMEN / SUMMARY

El objetivo es elaborar un marco teórico sobre los métodos multivariados y realizar una aplicación, presentando los elementos necesarios para el análisis de información cualitativa y cuantitativa. Se utilizó para la aplicación información de los desocupados del PGP proveniente de la EPH. De los resultados se concluye la importancia de la utilización de estos métodos para el análisis de problemas sociales. Se considera que el trabajo resulta útil porque: 1.- se realiza una presentación esquematizada de los métodos multivariados en general y del análisis de componentes principales y *cluster* en particular; 2.- puede servir de base para la realización de otros trabajos, no necesariamente sobre la misma temática, basados en la estructura de análisis desarrollada; y 3.- se presenta una caracterización de los desocupados del PGP agrupados en tres tipologías, de acuerdo a su situación socioeconómica.

The objective of this paper is to elaborate a framework on a multivariate analysis, and an application presenting the key elements to analyse qualitative and quantitative data. To this application data about unemployed people of the General Pueyrredon District were used. We concluded on the significance of this kind of methods to analyse social problems. It is considered that the paper is useful because: 1.- It gives a presentation of the multivariate statistical methods, in general and of principal component and cluster analysis, in particular; 2.- this work could be useful to do other works over this analysis structure; and 3.- it provides a characterisation of unemployment in the General Pueyrredon District.

PALABRAS CLAVE / KEY WORDS

Análisis estadístico multivariado – análisis por clasificación - análisis de componentes principales – desocupación.

Multivariate statistical methods – cluster analysis - principal component analysis - unemployment.

* Universidad Nacional de Mar del Plata.
** Universidad Nacional de General Sarmiento.

1. Introducción

Este trabajo tiene por objetivo realizar una aplicación de dos métodos estadísticos multivariados (análisis de componentes principales y análisis de *cluster*) sobre la información de desocupación captada por la EPH en la ciudad de Mar del Plata en octubre de 1995.

Estos permiten observar cuáles son las características que diferencian a los individuos entre ellos y la realización de agrupamientos en función de las más relevantes.

El Instituto Nacional de Estadísticas y Censos (INDEC) desarrolla desde 1974 la Encuesta Permanente de Hogares (EPH), relevando información sobre distintos aspectos económicos y sociales. La EPH es realizada en dos ondas anuales (mayo y octubre). En octubre de 1995 se incorporaron al relevamiento nuevos aglomerados, entre los que se encuentra Mar del Plata. El resultado para dicha ciudad mostró uno de los mayores niveles de desocupación del país (22,1%). La relevancia de este hecho justifica la realización del trabajo.

Los resultados del trabajo muestran que de acuerdo a las características de los desocupados del PGP se pueden conformar tres agrupamientos. Dos de ellos están conformados por individuos que tendrían una situación socioeconómica relativamente buena, debido a que su problemática estaría contenida en el seno de sus familias. Mientras que un tercer grupo, que es el más numeroso (65% de los desocupados), tendría una situación socioeconómica más comprometida, al pertenecer a familias con bajos ingresos y pocos miembros ocupados y percibiendo ingresos.

También se encontró que el tamaño del grupo familiar al que pertenece el desocupado tiene un efecto atenuante de su problema socioeconómico. Así como el nivel educativo de la familia tiene una relación directa sobre la situación socioeconómica familiar.

El trabajo está estructurado de esta manera: en el apartado siguiente se desarrollará el marco teórico, referente a los métodos estadísticos multivariados, basado en los libros de Manly (1988) y de Lebart y otros (1995). En el capítulo tres, se avanza en la aplicación de estos métodos sobre la información de los desocupados del PGP y, en el último capítulo, se presentan las reflexiones finales.

2. Marco Teórico

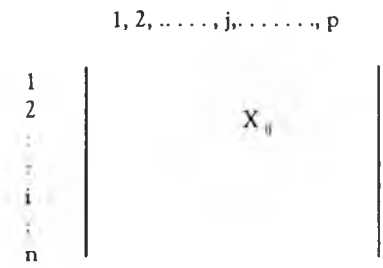
En este apartado se presentan los métodos multivariados en general, las dos subdivisiones en particular sobre las que se trabajará —análisis de componentes principales y *cluster*— y por último la complementariedad existente entre los métodos factoriales y los de clasificación.

Los métodos univariados (MU) analizan las variaciones ocurridas en una única variable aleatoria. Un ejemplo de esto puede ser la regresión múltiple que intenta explicar la variación de una variable dependiente a partir de 'n' variables independientes.

Los métodos multivariados (MM), en cambio, consideran un conjunto de variables aleatorias relacionadas, donde cada una es considerada igualmente importante al comienzo del análisis. Estos métodos significan un avance con relación a los MU, ya que permiten considerar el conjunto de variables simultáneamente, para luego intentar determinar cuáles son sus relaciones, en vez de establecer a priori las posibles relaciones entre las variables.

Si bien los MM se conocen desde comienzo de siglo, su utilización en forma generalizada en los últimos años ha sido posible gracias a los avances de la microelectrónica, que permitieron la utilización de programas estadísticos sofisticados en forma amplia.

En general, los MM utilizan una matriz de datos para organizar la información, y ésta tiene la forma:



donde las filas son los individuos que van desde i = 1... n, y las columnas las variables que van desde j = 1.... p. Cada celda de la matriz, cuando las variables son cuantitativas, mide el atributo o variable Xj para el individuo iésimo.

Los MM pueden dividirse en dos grandes grupos:

Los *de reducción* (métodos factoriales): intentan disminuir la dimensión del análisis reduciendo el número de variables necesarias a considerar a una cantidad menor, que igualmente capte una gran proporción de la variabilidad de los datos. Dentro de estos métodos se encuentran el análisis de componentes principales (ACP) y el análisis factorial (AF).

Los *de agrupamiento* (métodos de clasificación): intentan formar grupos de variables sobre la base de las medidas disponibles. Dentro de estos métodos se encuentran el análisis de la función discriminante (AFD) y el análisis de *cluster* (AC).

A continuación se realiza una breve descripción de estos métodos.

ACP: está diseñado para reducir el número de variables necesarias a considerar a un número de índices, llamados componentes principales, que son combinaciones lineales de las variables originales. Se utiliza con datos cuantitativos, aunque pueden ponerse como variables complementarias los cualitativos.

AF: también intenta explicar la variación en un número original de variables, usando un menor número de éstas indexadas, llamadas factores. Se asume que cada variable original puede ser expresada como una combinación lineal de estos factores más un término residual, que refleja el grado en el cual la variable es independiente de otras variables. Algunos análisis factoriales usan el ACP para encontrar los factores a utilizar.

AFD: busca separar diferentes grupos de variables sobre la base de las medidas disponibles. Como el ACP, el AFD está basado en la idea de encontrar combinaciones lineales adecuadas para las variables originales. Generalmente se trabaja sobre la base de 'm' muestras aleatorias de diferentes grupos, de tamaños $n_1, n_2, \ldots n_m$, y los valores estarán disponibles para 'p' variables, en cada miembro de la muestra. Se trata de ver el grado en que es posible discriminar entre grupos de individuos, sobre la base de combinaciones de variables a priori, partiendo de divisiones conocidas.

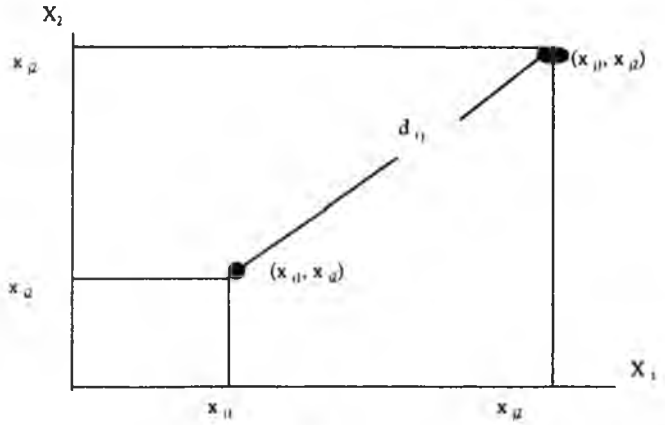
AC: intenta identificar grupos de individuos similares. Los grupos en este caso no son conocidos a priori sino que se trata de armarlos sobre la base de algún criterio, utilizando las variables disponibles.

2.1. Análisis de Componentes Principales

El ACP trabaja sobre la matriz de datos original, intentando reducir su dimensión para un mejor entendimiento de los datos a partir de seleccionar un menor número de variables. Busca alguna dirección o vector, tal que si proyectamos todos los individuos en esa dirección se

minimice la suma de los cuadrados de las diferencias o sea se minimice la distancia desde los distintos puntos al vector.

La distancia euclidiana entre dos variables (d_{ij}), para el caso de p = 2, puede calcularse utilizando el teorema de Pitágoras, haciendo la raíz cuadrada de la suma de los cuadrados de las diferencias. Por ejemplo, para el caso:



la distancia puede calcularse como: d_{ij} = Raíz² í($x_{i1} - x_{j1}$) ² + ($x_{i2} - x_{j2}$) ²ý.

Generalizando, si consideramos 'p' variables la distancia euclidiana entre dos puntos (i, j) puede calcularse como:

$$d_{ij} = \text{Raíz}^2 \{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \}$$

El ACP intenta reducir un gran número de variables originales a un pequeño número de componentes principales, buscando un vector Z con 'p' componentes, una nueva dimensión, tal que si se proyectan todas las dimensiones originales sobre éste se minimicen las distancias euclidianas con respecto a las variables originales.

Cada componente del vector es una combinación lineal de las variables originales. El objetivo de este método es tomar 'p' variables X_1, X_2, ..., Xp y encontrar combinaciones de éstas produciendo los índices Z_1, Z_2, ..., Z_p que no están correlacionados. La falta de correlación es una propiedad útil, dado que esto significa que los índices están midiendo diferentes 'dimensiones' en los datos. Sin embargo, los índices están también ordenados de forma tal que Z_1 capta la variación de mayor tamaño, Z_2 muestra la segunda mayor variación, y así sucesivamente, en orden decreciente. Así, var(Z_1) ³ var (Z_2) ³ .... ³ var (Z_p), donde var (Z_j) denota la varianza de Z_j en el conjunto de datos considerados. Z_1 es el llamado componente principal. Cuando se realiza un análisis de componente principal, para lograr el objetivo de reducción, se requiere que los primeros componentes expliquen una gran parte de la variabilidad de los datos. O, lo que es lo mismo, se espera que las varianzas de la mayoría de los índices sean insignificantes. En tal caso, las variaciones en los datos pueden ser adecuadamente descriptas por un pequeño número de variables Z con varianzas que no son insignificantes. De esta forma se logra un cierto grado de economía, dado que de las 'p' variables originales X se arriba a un pequeño número de variables Z.

Así, si partimos de 'p' variables y 'n' individuos, el primer componente principal (Z_1) será la combinación lineal de las variables originales X_1, X_2, ..., X_p

$$Z_1 = a_{11} X_1 + a_{12} X_2 + \ldots\ldots + a_{1p} X_p$$

que varía tanto como sea posible para los individuos, sujeto a la condición de que:

$$a_{11}^2 + a_{12}^2 + \ldots\ldots + a_{1p}^2 = 1.$$

Luego, la varianza de Z_1, var (Z_1) puede ser tan grande como sea posible, dado la restricción sobre la constantes a_{1j}. La restricción es introducida porque si no la var (Z_1) puede ser incrementada simplemente incrementando cualquiera de los a_{1j} valores.

El proceso es igual para los 'p' componentes principales. Luego deben elegirse aquellos componentes que expliquen una proporción significativa de la varianza total de los datos.

Para encontrar los a_{ij} el procedimiento es el siguiente:

Se suele iniciar el análisis estandarizando los valores de las variables originales para que tengan media (C) = 0 y varianza $(s^2) = 1$. Esto se realiza para evitar que una de las variables tenga una influencia indebida sobre el componente principal.

Se calcula la matriz de covarianzas sobre los valores estandarizados (que es la misma que la de correlaciones para las variables estandarizadas).

$$C = \begin{vmatrix} c_{11}\, c_{12} \ldots\ldots\ldots\ldots\ldots c_{2p} \\ c_{21}\, c_{22} \ldots\ldots\ldots\ldots\ldots c_{2p} \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ c_{p1}\, c_{p2} \ldots\ldots\ldots\ldots\ldots c_{pp} \end{vmatrix} = \begin{vmatrix} 1 \;\; r_{12} \ldots\ldots\ldots\ldots\ldots r_{1p} \\ r_{21}\; 1 \ldots\ldots\ldots\ldots\ldots r_{2p} \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ r_{p1}\, r_{p2} \ldots\ldots\ldots\ldots\ldots 1 \end{vmatrix}$$

Donde $c_{ii} = 1$ debido a que es igual a la varianza (s^2) de las variables estandarizadas y $c_{ij} =$ cov (X_i, X_j) es igual a $r_{ij} =$ cov $(X_i, X_j) / s^2_i\, s^2_j$, donde ambos componentes del denominador son iguales a 1. A su vez, $c_{ij} = c_{ji}$ o $r_{ij} = r_{ji}$, por lo que la matriz es simétrica.

3. Se deben encontrar los valores característicos de la matriz C; $d_1\ d_2 \ldots\ldots d_p$ y los correspondientes vectores característicos $A_1, A_2, \ldots, A_p$ cuyos componentes serán los valores que se utilizan para construir las combinaciones lineales en los Z_i

$$A_1 = \begin{vmatrix} a_{11} \\ a_{12} \\ \vdots \\ \vdots \\ \vdots \\ a_{1p} \end{vmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad A_p = \begin{vmatrix} a_{p1} \\ a_{p2} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ a_{pp} \end{vmatrix}$$

Por lo tanto, los coeficientes para el iésimo componente principal son los a_i mientras que d_i es su varianza.

4. Se deben descartar los componentes principales que sólo expliquen una pequeña proporción de la varianza en los datos y de esta forma se reducen las dimensiones del análisis. O, lo que es lo mismo, para reducir la dimensión del análisis se eligen aquellos componentes principales cuyas varianzas expliquen la mayor proporción de la varianza total, utilizando mayores (di/p).

Para determinar con cuántos componentes trabajar puede utilizarse alguna de las siguientes reglas de corte:

a. Regla del salto (regla del codo): observando el histograma de los valores propios, la existencia de un salto importante en las barras muestra donde puede realizarse el corte.

b. Promedio de los valores propios: tomo todos los componentes que su valor propio supera o es igual al promedio de los valores propios.

En resumen, las varianzas de los componentes principales son los valores característicos de la matriz C (de varianzas y covarianzas o de correlaciones). Si los valores característicos son ordenados tal que $d_1 = d_2 = \ldots\ldots = d_p = 0$, luego el d_i es la varianza del iésimo componente principal y el vector A_i proporciona los coeficientes para construir la combinación lineal de las variables originales que lo forman.

Una propiedad importante de los valores característicos es que:

$d_1 + d_2 + \ldots\ldots + d_p = c_{11} + c_{22} + \ldots\ldots + c_{pp}$ y en el caso de que las variables estén estandarizadas esta suma es igual a 'p' ya que los $c_{ii} = 1$.

Además, como los c_{ii} son las varianzas de las variables X_i y los d_i las varianzas de los Z_i, la suma de las varianzas de los componentes principales es igual a la suma de las varianzas de las variables originales. Esto quiere decir que los componentes principales explican todas las variaciones en los datos originales.

Por otra parte, para reducir la dimensión del análisis se eligen aquellos componentes principales cuyas varianzas expliquen la mayor proporción de la varianza total, utilizando mayores (d_i/p). Para que el procedimiento tenga sentido se necesita que las variables estén altamente correlacionadas, y de esta manera algunas puedan eliminarse. De otra manera todas son importantes para explicar la varianza total y no podría eliminarse ninguna.

2.2. Análisis de *cluster*

Según Manly (1986), el análisis de *cluster* (de clasificación o de clasificación automática) es utilizado para solucionar el problema de asignar aquellos objetos (individuos) con características "similares" a un determinado grupo. El método debe ser completamente numérico y el número de grupos es desconocido.

Los datos para un análisis de *cluster* usualmente consisten en 'p' variables $X_1, X_2, \ldots\ldots, Xp$ por 'n' objetos. Los valores de las variables son usados para producir luego una serie de distancias entre los individuos. Las medidas de las distancias, como por ejemplo la función de distancia euclidiana, ya fueron tratadas en la sección 2.1.

Usualmente, las variables están estandarizadas en cierta forma antes de ser calculadas las distancias, por lo tanto todas las 'p' variables son igualmente importantes en la determinación de dichas distancias (medias iguales a cero y varianzas iguales a uno). Alternativamente, cada variable puede ser recalculada para tener un mínimo igual a cero y un máximo de uno.

Existen dos grandes familias de métodos estadísticos que permiten clasificar un conjunto de unidades de observación: el enfoque jerárquico y el de centros móviles.

El primero de los métodos comienza con el cálculo de las distancias de cada individuo hacia los otros. Luego son formados grupos mediante un proceso de aglomeración o división. Con la aglomeración todos los objetos comienzan estando solos, en grupos de uno. Grupos "cerrados" son luego gradualmente unidos hasta que finalmente todos los individuos están en un único grupo. Con el proceso de división todos los objetos comienzan en un único grupo. Este es luego dividido en dos grupos, los dos grupos son luego divididos, y así sucesivamente hasta que los objetos terminan estando solos ("grupos" de un individuo). En cualquiera de

estos métodos uno debe decidir cuándo detener el proceso de agrupamiento o división.

El segundo enfoque para el análisis de *cluster* comienza con una decisión a priori de cuántos grupos se van a formar y con la elección, más o menos arbitraria, del centro de los grupos y los individuos, que serán asignados de acuerdo a su proximidad con cada uno de estos centros. Nuevos centros son luego calculados y los individuos son movidos a un nuevo grupo si está más cerca del nuevo centro que del anterior. Grupos completos pueden ser fusionados al nuevo centro; otros grupos pueden ser divididos, etc. El proceso continúa hasta que la estabilidad es lograda con el número predeterminado de grupos.

Así, la diferencia entre los dos enfoques reside en la forma en que se calculan las distancias. A continuación se desarrolla con mayor profundidad el método jerárquico.

2.2.1. Método jerárquico

Las clasificaciones jerárquicas tienen como objeto presentar de manera sintética el resultado de las comparaciones entre objetos de una tabla observada (individuos, modalidades o variables).

Una clasificación jerárquica es una serie de particiones encajadas.

Como ya se mencionó, el método jerárquico de aglomeración comienza con una matriz de 'distancias' entre individuos. Todos los individuos empiezan estando solos, en grupos de uno, y luego los grupos cercanos son fusionados. Por ejemplo, supongamos la siguiente matriz para cinco objetos:

	1	2	3	4	5
1	-				
2	2	-			
3	6	5	-		
4	10	9	4	-	
5	9	8	5	3	-

Los cálculos son luego realizados como se muestra en la tabla siguiente. Los grupos son fusionados a un nivel dado de distancia si uno de los individuos en un grupo se encuentra cerca de, al menos, un individuo del segundo grupo.

Distancia	Grupos
1	1, 2, 3, 4, 5
2	(1,2), 3, 4, 5
3	(1,2), 3, (4,5)
4	(1,2), (3,4,5)
5	(1,2,3,4,5)

A una distancia de 0 los cinco objetos están en forma independiente. La matriz de distancias muestra que la menor distancia entre dos elementos es 2, entre el primero y segundo elemento. Así, a una distancia de 2 hay 4 grupos (1,2), (3), (4) y (5). La próxima menor distancia entre objetos es 3, entre los objetos 4 y 5. Por lo tanto, a una distancia de 3 hay tres grupos (1,2). (3) y (4,5). La próxima menor distancia es 4, entre los objetos 3 y 4. Por lo tanto,

a esta distancia quedan conformados dos grupos (1,2) y (3,4,5). Finalmente, la próxima menor distancia es 5, entre los elementos 2 y 3 y entre los elementos 3 y 5. A esta distancia los dos grupos se juntan en un único grupo (1,2,3,4,5) y el análisis es completo.

Por lo tanto, las etapas fundamentales del método jerárquico pueden presentarse de la siguiente forma:

Etapa 1: hay 'n' elementos a clasificar (que son los 'n' individuos);

Etapa 2: se construye la matriz de distancias entre los 'n' elementos y luego se busca los dos más próximos, que se constituyen en un nuevo elemento. Obteniéndose una primera partición en *n-1* clases;

Etapa 3: se construye una nueva matriz de distancias entre el nuevo elemento y los elementos restantes (las otras distancias son inalteradas). Otra vez nos encontramos en las mismas condiciones que en la etapa 1, pero con solamente *n-1* elementos a clasificar. Nuevamente se buscan los dos elementos más próximos, que luego son agrupados. Obteniéndose una segunda partición con *n-2* clases y que engloban a la primera.

Etapa *m*: se calculan las nuevas distancias, y luego se reitera el proceso hasta reagrupar todos en un solo grupo, que constituye la última partición.

La representación gráfica del resultado de las comparaciones entre los individuos observados es llamada árbol de clasificación o dendrograma. En él se pueden ver los reagrupamientos sucesivos (figura del Anexo II).

A partir de un "árbol de clasificación" se puede elegir una "buena" partición de los "n" objetos sometidos a la clasificación jerárquica ascendente. Para ello, es suficiente "cortar" el dendrograma con una recta horizontal que cruce las ramas ascendentes más largas.

2.2.2. Problemas del análisis de cluster

Como ya se mencionó, existen varias formas para realizar el análisis de *cluster*. Sin embargo, no hay ninguna aceptada como la mejor. Desafortunadamente, diferentes métodos no necesariamente producen los mismos resultados a partir de un mismo conjunto de datos. Es usual que haya una gran dosis de subjetividad en la valoración de los resultados desde un método particular.

También es subjetiva la selección de variables, y el problema aquí reside en que los grupos obtenidos pueden ser sensitivos a un grupo particular de variables seleccionadas. Una elección distinta de variables, en apariencia igualmente razonables, quizás produzca grupos bastante diferentes.

2.3. Complementariedad entre los métodos de análisis factorial y de clasificación

Los métodos factoriales se adaptan particularmente bien a la exploración de grandes tablas de datos individuales, resultando su producto sumamente útil para la investigación. Sin embargo, no es suficiente para formar una visión totalmente satisfactoria de la relación entre los datos. No sólo los resultados no se vinculan con una parte de la información, sino que ellos son a veces muy complejos para ser interpretados fácilmente.

Ante estas circunstancias, las técnicas de clasificación pueden completar y matizar los resultados del análisis factorial. La complementariedad entre análisis factorial y clasificación concierne la comprensión de la estructura de datos y la ayuda práctica en la fase de interpretación de los resultados.

2.2.3. Necesidad e insuficiencia de los métodos factoriales

La representación gráfica a través de los métodos factoriales presenta algunos inconvenientes:

1) Dificultad de interpretación

Son muy difíciles de interpretar los ejes o planos factoriales. Los planos (3 y 4), generados por los ejes factoriales 3 y 4, describen la proximidad a la que están los términos correctivos resultantes de las principales proximidades observadas sobre los dos primeros ejes. La interpretación de dicha proximidad es, por lo tanto, bastante delicada.

2) Comprensión excesiva y deformación

La visualización está limitada a dos, o en general a muy pocas dimensiones, de forma que el nombre de los ejes "significativos" sea considerado un bien superior. Esta comprensión excesiva del espacio quizás sufra de la distorsión y superposición de los puntos que ocupan las distintas posiciones dentro del espacio.

3) Falta de robustez

La visualización puede carecer de robustez. Un dato erróneo puede notablemente influenciar al primer factor y a todas las dimensiones siguientes, dado que éstas son relativas al primer eje.

4) Gráfico factorial ilegible

La gráfica puede estar dada por centenares de puntos, resultando una visualización cargada o ilegible.

Para remediar estos problemas deberían utilizarse simultáneamente los aportes de un método de clasificación.

Dificultad de interpretación y comprensión excesiva de los datos (puntos 1 y 2)

Para completar el análisis factorial puede utilizarse una clasificación realizada sobre el total del espacio o sobre un sub-espacio definido por los primeros factores —los más significativos—. Las clases toman en cuenta la dimensión relativa de la nube de puntos. Ellas corrigen ciertas deformaciones debidas a la operación de proyección.

Una clase puede ser también típica de un eje de rango elevado y ayudar a la interpretación de ese sub-espacio en particular, difícilmente observable de otra manera.

Robustez imperfecta (punto 3)

La mayor parte de los algoritmos de clasificación, y particularmente los algoritmos de aglomeración, son robustos en el sentido de que los nodos correspondientes a las distancias más pequeñas son independientes de eventuales puntos individuales marginales.

Facilitación y descripción automática de los resultados gráficos (punto 4)

Dada la existencia de un conjunto de puntos individuales sobre el plano factorial, resulta útil proceder al reagrupamiento de los individuos en familias homogéneas. Las clases resultantes pueden ser utilizadas para ayudar a la interpretación de los planos factoriales, identificando las zonas que están bien descriptas. Es en efecto más fácil de describir estas clases que un espacio continuo. La noción de clase es elemental y accesible a la intuición. La descripción de estas clases puede estar fundada sobre la elemental comparación de medias y porcentajes. Los nombres de los puntos son así reemplazados por los centros de las distintas clases. Como los algoritmos matemáticos utilizados por este reagrupamiento funcionan de la misma forma que los puntos situados dentro de un espacio factorial de dos o de diez dimensiones, se puede decir que se mejora la calidad de representación.

"Aunque insuficientes, los métodos factoriales son necesarios: la facultad descriptiva de los ejes, la descripción bajo la forma de continuo geométrico resultan irremplazables." (Lebart y otros, 1995: p.187)

La clasificación no siempre tiene éxito al mostrar la importancia de ciertas tendencias o de factores latentes continuos. Para observar la organización espacial de las clases, el posicionamiento de éstas sobre los ejes factoriales resulta indispensable. La clasificación puede ayudar a descubrir la existencia de grupos de individuos. El análisis factorial pone en evidencia factores

latentes no atendidos. El descubrimiento de tales fenómenos o dimensiones escondidas es el objetivo de estas dos familias de métodos, y ciertamente la más ambiciosa. Por lo tanto, la utilización complementaria resulta indispensable para atender a ese objetivo.

A continuación se realiza la presentación y análisis de la información sobre los desocupados aplicando primero los métodos de ACP y AC en forma individual y luego en forma conjunta.

3. Estudio de la desocupación en Mar del Plata

3.1. Población y características de la muestra

Para este trabajo se utilizan datos de la Encuesta Permanente de Hogares (EPH) realizada en la ciudad de Mar del Plata, en octubre de 1995. La EPH es realizada por el Instituto Nacional de Estadística y Censo (INDEC) desde el año 1974, y tiene por objetivo llevar a cabo investigaciones continuas de diversos aspectos económicos y sociales, sobre una muestra de hogares urbanos de distintas partes del país.

El INDEC optó para la muestra por un diseño probabilístico bietápico, en el que se adoptaron como unidades de selección para la primera etapa los radios censales, grupos de radios o subdivisiones de ellos, según el caso, y para la segunda etapa, las viviendas (INDEC, 1995).

La fuente de datos sobre la que el INDEC realiza el diseño de la muestra actualmente es el Censo de Población de 1991. La EPH releva 28 aglomerados urbanos de todo el país, de los cuales cuatro pertenecen a la provincia de Buenos Aires y ellos son: Conurbano, Gran La Plata, Bahía Blanca y Mar del Plata.

Los criterios de elección del tamaño de la muestra también están dados por el Censo de1991, en donde a todas aquellas ciudades que tengan entre 80 mil y 300 mil viviendas les corresponde un tamaño de muestra igual a 800. Finalmente, y producto de algunos factores de corrección para Mar del Plata, se seleccionaron 822 viviendas.

Los campos que abarca esta encuesta son:

1. Características demográficas
2. Características ocupacionales
3. Características migratorias
4. Situación habitacional
5. Características educacionales
6. Distribución de ingresos

La captación de la información se realiza a través de dos tipos de cuestionarios, uno para el grupo familiar y otro por individuo. Los atributos que se quieren obtener de la población son inherentes a los individuos, por lo tanto éste constituye la unidad elemental o de análisis, siendo la vivienda la unidad entrevistada.

3.2. Aplicación del Análisis de Componentes Principales

Tal como se indica en el apartado 3.1, la cantidad de viviendas sobre las que trabaja la EPH en la ciudad de Mar del Plata asciende a 822. *Para esta aplicación se utiliza sólo la información sobre los 186 individuos de la muestra que se encontraban desocupados, en la onda de octubre de 1995*, sobre un total de 819 individuos que componen la población económicamente activa[1] en la muestra.

En este apartado, se analizan los individuos y las variables utilizando el programa SPADN versión 2.52. La matriz de datos "teórica" sobre la que se trabaja es la siguiente:

[1] Población Económicamente Activa (PEA): la integran las personas que tienen una ocupación o que, sin tenerla, buscan activamente. Está compuesta por la población ocupada más la población desocupada (Lacabana y otros, 1997).

Individuos	VARIABLES ACTIVAS (15)										VARIABLES ILUSTRATIVAS (10)							
	A	B	C	D	E	F	G	H	-	AABW	FALT	TEMP	ULTI	VALE	LEER	ESCU		CIVI
1																		
2																		
:																		
:																		
:																		
:																		
:																		
:																		
186																		

3.2.1. Análisis de los individuos

Si observamos los individuos, su distribución y densidad (figura 1), obtenemos una primera información sobre cómo son determinados los componentes principales. Los valores más distantes del centro del eje de coordenadas y más cercano a cada uno de los ejes (con menor ángulo) son los que explican un porcentaje mayor de la varianza de cada eje (en la figura 1 puede observárselos encerrados en círculos). Si todos los valores de la nube de puntos hubieran estado concentrados la aplicación de este método no hubiera tenido sentido.

3.2.2. Análisis de las variables activas

En primer lugar, se presentan los parámetros que mejor describen la muestra. Luego, se utiliza el análisis de componentes principales para reducir la dimensión del espacio y poder determinar cuáles son las principales variables que explican la variabilidad de la muestra.

La lectura del cuadro 1 permite una primera caracterización de la información sobre los desocupados de Mar del Plata (la tabla de variables se presenta en el Anexo I). Por ejemplo, se observa que cada vivienda tiene en promedio un ocupado, la habitan cuatro personas y uno de ellos percibe ingresos. Sin embargo, estas medidas deben analizarse teniendo en cuenta los valores de la desviación estándar y los valores extremos, dado que, por ejemplo, si bien el ingreso total de la familia en promedio es de 700 pesos, hay familias que perciben cero pesos y otras casi cuatro mil.

Cuadro 1: Estadísticos básicos

Variable	*Media*	*Desv. Estánd.*	*Mínimo*	*Máximo*
A	7.91	10.68	0.30	72.0
B	21.41	40.90	0.00	300.0
C	49.70	115.35	0.00	600.0
D	92.19	193.67	0.00	1500.0
E	8.71	3.25	0.00	18.0
F	22.32	14.75	1.00	65.0
G	699.78	663.48	0.00	3954.0
H	167.10	157.49	0.00	833.3
I	35.44	15.25	13.00	81.0
J	3.05	1.39	1.00	9.0
K	4.54	2.48	1.00	13.0
L	1.07	1.06	0.00	6.0
M	1.44	0.64	1.00	3.0
N	2.03	1.85	0.00	8.0
O	1.67	1.24	0.00	7.0

Siguiendo los pasos para el análisis de componentes principales, a continuación se expone y analiza la matriz de correlaciones. Su lectura nos da una primera idea del grado de interrelación existente entre las variables y, en la medida que éste sea elevado, la aplicación del análisis de componentes principales nos permitirá sintetizar la información. La importancia de la matriz radica en: 1) si los valores observados de correlación son altos, la posibilidad de reducir las dimensiones del espacio es muy alta y viceversa, y 2) si los valores no son particularmente altos, esto indicaría que varios componentes principales serán requeridos para captar las variaciones existentes.

Cuadro 2: Matriz de Correlación

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A	1														
B	.27	1													
C	.21	.11	1												
D	-.07	.00	.19	1											
E	.13	-.04	-.09	.06	1										
F	.12	.17	.24	-.09	-.21	1									
G	-.01	.04	-.02	.09	.22	-.19	1								
H	.07	.06	.07	.27	.33	-.12	.72	1							
I	.09	.22	.17	.13	-.20	.51	-.26	-.04	1						
J	.02	.07	-.06	.01	.29	-.13	.37	.20	-.15	1					
K	-.01	.04	.02	-.08	-.18	-.16	.41	-.15	-.35	.24	1				
L	-.01	-.13	-.05	-.21	-.04	.14	.60	.17	-.30	.19	.60	1			
M	.02	.09	.07	.00	-.16	.02	-.01	-.18	-.09	.12	.37	.08	1		
N	-.13	.10	.03	.01	-.16	-.14	.21	-.24	-.27	.17	.86	.21	.11	1	
O	-.02	.01	.03	.11	-.02	-.08	.64	.25	-.19	.32	.63	.70	.25	.36	1

Los valores en bastardilla en el cuadro 2 son los que tienen correlaciones superiores en valor absoluto a 0.4.

Las variables A, B, C, D, E, J y M registran en general bajos valores de correlación con el resto de las variables.

El tiempo de residencia en Mar del Plata (F) está correlacionado positivamente con la edad del desocupado (I).

El ingreso total de la familia (G) está relacionado positivamente con el ingreso promedio por componente familiar (H), con la cantidad de personas que habitan en la vivienda (K), con la cantidad de ocupados que habitan en la vivienda (L) y con la cantidad de personas que perciben ingresos en la familia (O).

La cantidad de personas que habitan en la vivienda (K) está positivamente relacionada con la cantidad de ocupados que habitan en la vivienda (L), con la cantidad de personas inactivas que habitan la vivienda (N) y con la cantidad de personas que perciben ingresos en la vivienda (O).

La cantidad de ocupados que habitan en la vivienda (L) está positivamente correlacionada con las personas que perciben ingresos y habitan la vivienda.

Entonces, dada la existencia de variables altamente correlacionadas, es posible la construcción de algunos componentes principales que expliquen un porcentaje significativo de la variación en los datos.

El cuadro 3 resume algunas de las características de los componentes principales que surgen a partir de las variables consideradas —los valores propios y los porcentajes de varianza—. La columna 2 muestra los valores propios, éstos son la varianza de cada componente principal. Su suma es igual a la traza de la matriz C y coincide con el número de variables (15). El primero de los valores propios, el de mayor valor, explica el 24% de la varianza total de los datos estandarizados. Al mismo tiempo, el vector característico A_1 asociado a este valor propio proporciona los coeficientes del primer componente principal (Ver cuadro 4). El segundo valor propio explica el 15% de la varianza y su vector asociado A_2 proporciona los coeficientes para el segundo componente principal.

Cuadro 3: Histograma de los valores propios

Nr	Valor Propio	%	% acum	
1	3.6500	24.33	24.33	•••
2	2.2058	14.71	39.04	••
3	1.8095	12.06	51.10	••
4	1.2455	8.30	59.41	••••••••••••••••••••••••••••
5	1.1808	7.87	67.28	••••••••••••••••••••••••••
6	0.9417	6.28	73.56	•••••••••••••••••••••
7	0.9265	6.18	79.73	•••••••••••••••••••••
8	0.7914	5.28	85.01	••••••••••••••••••
9	0.6080	4.05	89.06	••••••••••••••
10	0.5365	3.58	92.64	••••••••••••
11	0.4431	2.95	95.59	••••••••••
12	0.3695	2.46	98.06	•••••••••
13	0.2040	1.36	99.42	•••••
14	0.0877	0.58	100.00	••
15	0.0000	0.00	100.00	•

Entonces, si se toman los dos primeros componentes principales, el porcentaje de la varianza explicado asciende a 39,04%, viéndose en el histograma del cuadro 3 que existe una concentración dentro del sub-espacio de dos dimensiones. La observación de los "saltos" en el histograma proporciona un criterio para la elección del número de componentes a seleccionar. En este caso, se observa un gran salto entre el primero y segundo, así como entre el segundo y tercero, por lo que se podrían seleccionar los dos primeros o los cuatro primeros. La alternativa elegida en este trabajo es la primera, y se fundamenta en la magnitud del salto, en el porcentaje de varianza explicado y en la facilidad que esto proporciona para el análisis gráfico; asumiendo que es requerido un pequeño número de índices para presentar los principales aspectos que caracterizan a los individuos de la muestra.

Finalmente, es interesante señalar que si los saltos observados en el histograma fueran de pequeña magnitud, no se podría reducir las dimensiones del espacio original. Esto sucedería en el caso de una baja correlación entre las variables originales.

Cuadro 4: Vectores característicos (Coef. de los componentes principales)

Variable	A1	A2
A	.14	.18
B	.05	-.02
C	.08	-.03
D	.01	.28
E	-.09	.64
F	.36	-.20
G	-.76	.48
H	-.28	.82
I	.52	-.02
J	-.46	.25

Continúa

K	-.83	-.50
L	-.77	-.02
M	-.24	-.39
N	-.59	-.52
O	-.82	.01

En el cuadro 4 se presentan los coeficientes de los dos primeros componentes principales. En donde el primer componente principal es:

$$Z_1 = 0.14\ A + 0.05\ B + 0.08\ C + 0.01\ D - 0.09\ E + \ldots\ldots - 0.82\ O$$

Las variables en esta ecuación están estandarizadas, teniendo media igual a cero y desvío estándar igual a uno. Para el análisis, las variables con coeficientes cercanos a cero son ignoradas dado que ellas no afectan al valor de Z_1 en forma significativa.

Si observamos los coeficientes del primer componente podemos ver, en primera instancia, un contraste debido a su signo, entre las variables F (tiempo que hace que vive en Mar del Plata) e I (edad del individuo) y los de las variables G (ingreso total de la familia), K (cantidad de personas que habitan en la vivienda) y O (cantidad de personas que perciben ingresos en la vivienda).

El segundo componente viene determinado por la ecuación:

$$Z_2 = 0.18\ A - 0.02\ B - 0{,}03\ C + 0{,}28\ D + 0{,}64\ E + \ldots\ldots - 0{,}01\ O$$

Aquí encontramos un contraste entre los coeficientes de la variables E (años de educación) y H (ingreso promedio por componente principal) y los de la variable K (cantidad de personas que habitan en la vivienda), M (cantidad de desocupados que habitan la vivienda) y N (cantidad de personas inactivas que habitan la vivienda).

La figura 2 (Anexo IV) muestra las distintas variables sobre los dos primeros ejes factoriales. Los datos están aquí centrados (reducidos) y las coordenadas de las variables sobre los ejes —coeficientes de los componentes principales— son los valores de correlación entre las variables y los factores.

Así, el primer eje contrapone *variables relacionadas con el transcurso del tiempo* (edad del individuo, tiempo que hace que reside en Mar del Plata) con *variables que tienen que ver con el tamaño del grupo familiar y sus ingresos* (cantidad de ocupados que habitan la vivienda, cantidad de personas que perciben ingresos e ingreso total de la familia). Por otro lado, el segundo eje contrapone *variables que tienen que ver con el nivel educativo e ingresos del individuo* (años de educación e ingreso promedio por componente familiar) con *variables que tienen que ver con el número de desocupados por vivienda* (cantidad de desocupados que habitan la vivienda y cantidad de personas inactivas que habitan la vivienda).

La figura 2 presenta también las variables ilustrativas. Para el presente análisis las variables seleccionadas aportan información que parecería ser no muy contundente dada su cercanía al eje de coordenadas. No obstante, en el eje 2 la variable LEER (sabe leer y escribir) podría estar apoyando la idea de que ese eje está relacionado con el nivel educativo, y la variable TEMP (trabajo temporario) podría estar relacionada con el número de desocupados por vivienda. A su vez, en el eje 1 la variable CIVI (estado civil) podría estar relacionada con la edad de los individuos.

A continuación se complementará el estudio mediante la utilización del análisis de *cluster*.

3.3. Aplicación del análisis de *cluster*

La aplicación del método de análisis de *cluster* es realizada bajo el enfoque jerárquico de aglomeración, aplicando el programa SPADN 2.52 y utilizando los datos resultantes del análisis de componente principal. Es decir, se trabaja con las mismas variables pero solucionando el problema de correlación en el que podría incurrirse si se utilizaran los valores originales.

3.3.1. Clasificación de los individuos

El análisis de *cluster* no trata de reducir el número de variables a analizar (como lo hace el ACP), sino que trata de juntar a los individuos en grupos, de acuerdo a sus características.

Los principales resultados de esta clasificación son resumidos en el cuadro presentado en el Anexo II. En éste se presenta la información de la clasificación jerárquica correspondiente a los 49 nodos con índices más elevados, la cual puede leerse de la siguiente forma: la primera columna (NUM) da los números de nodos, los cuales son los nuevos elementos a clasificar. La columna 2 y 3 (llamadas Primogénito y Benjamín) es aplicada a los dos elementos que son agrupados en una etapa dada (se puede decir los más próximos en cierta etapa en el sentido del índice de agregación retenido).

La primera línea de dicho cuadro presenta al nodo 323, que estando formado por los dos elementos más próximos 272 y 277 (Primogénito y Benjamín), conforman un nuevo grupo formado por 7 elementos (columna EFE), donde el peso total (columna PESO) es igual a 7. El valor del índice de agregación correspondiente es de 0,00285. Los valores crecientes del índice son ilustrados a través del histograma que se presenta al lado de las columnas numéricas (este histograma puede dar una idea del número de clases de una buena partición, si la distancia entre las dos clases más próximas es grande, al igual que el valor del índice es elevado, se observa un salto importante en el histograma, por lo que se puede esperar la obtención de una partición de buena calidad).

El árbol jerárquico del Anexo III da la misma información, pero presentada de una forma más ilustrativa, pues se mantiene visible la forma en que se van armando los grupos a partir de los elementos más próximos.

A partir de esta información se puede observar en el Anexo II que, tomando en cuenta los saltos en el histograma para la partición en clases, es importante la variación (salto) entre los índices de los nodos 371 y 370 —por lo que se podrían conformar dos grupos— y también la variación (salto) entre los índices del nodo 370 y 369 —por lo que se podrían conformar tres grupos—. Las variaciones del índice tienen algunos otros saltos visibles pero de una magnitud considerablemente menor. Para este trabajo se usará una partición en tres grupos. Esto permitirá realizar un análisis más amplio que el que se haría si sólo se trabajara con dos clases.

3.3.2. Descripción estadística de los grupos

La descripción automática de los grupos constituye en la práctica una etapa indispensable de todo proceso de clasificación. La idea de la interpretación de los grupos está generalmente fundada en la comparación de medias o de porcentajes al interior de cada grupo con las medias o los porcentajes obtenidos del total de los elementos a clasificar.

Para seleccionar las variables características se apela a un *valor-test*. Dicho valor $t_k(X)$ evalúa la distancia entre la media dentro del grupo ($\bar{X}_k$) y la media general ($\bar{X}$) y ajusta esta distancia tomando en cuenta el peso del desvío típico del grupo ($s_k(X)$). El *valor-test* es simplemente la cantidad:

$$t_k(X) = \frac{\bar{X}_k - \bar{X}}{s_k(X)}$$

donde:

$$s_k^2(X) = \frac{n - n_k}{n - 1} \frac{s^2(X)}{n_k}$$

Por lo tanto, los valores de $t_k(X)$ serán mayores cuanto más alto sea el valor de la media del grupo y menores sean los valores de la media general y del desvío estándar.

Las variables suplementarias, si bien no participan en la construcción de los grupos, son presentadas también dado que pueden ayudar a definir estas particiones.

En la información relativa al *grupo 1* (cuadro 5) figuran los datos de las variables activas e ilustrativas más significativas. Este grupo está compuesto por el mayor número de individuos (121, 65%) y caracterizado por una edad media de los componentes del grupo (39) relativamente superior a la edad media promedio general (35); por ingresos familiares ($385.85), número de receptores de ingresos en la familia (1), ocupados en la vivienda (0-1), ingresos por componente familiar ($108.81) y un tamaño familiar (3-4) inferior a la media general. Esto indica que el grupo 1 está compuesto por las familias más pequeñas y con ingresos sustancialmente menores, en las cuales la edad de los desocupados es en promedio más alta que la media del total de los desocupados. A su vez, son relativamente significativas (su valor test no es muy alto) las variables ilustrativas Estado Civil y Relación de Parentesco con el Jefe de Familia, lo cual indicaría que una proporción importante de los desocupados del grupo estarían casados y serían o jefe de familia o su cónyuge.

Cuadro 5: Caracterización del grupo 1

CLASE 1 / 3 (PESO = 121; EFECTIVO = 121) - IDENT. aala -							
V.TEST	PROB.	MEDIAS		DESVIO ESTANDAR		VARIABLES CARACTERISTICAS	
		CLASE	GENERAL	CLASE	GENERAL	NUM. ETIQUETA	IDEN
4,41	0,000	39,07	35,44	15,17	15,25	9. EDAD	I
3,78	0,000	2,45	2,22	1,10	1,13	43. ESTADO CIVIL	CIVI
3,75	0,000	25,31	22,32	16,09	14,75	6. ANTIGÜEDAD EN MAR DEL PLATA	F
2,65	0,004	1,95	1,90	0,25	0,33	37. ASISTIO A LA ESCUELA	ASIS
-2,92	0,002	8,20	8,71	2,97	3,25	5. AÑOS DE EDUCACION	E
-3,88	0,000	1,64	2,03	1,40	1,85	14. INCATIVOS	N
-4,74	0,000	1,76	2,00	0,79	0,94	16. PARA QUE BUSCA TRABAJO	SPQB
-4,79	0,000	2,69	3,05	1,08	1,36	10. HABITACIONES	J
-5,13	0,000	1,79	2,10	0,93	1,10	41. RELACIÓN	RELA
-6,24	0,000	3,71	4,54	1,42	2,48	11. PERSONAS EN LA VIVIENDA	K
-6,87	0,000	108,81	167,10	80,23	157,49	8. INGRESO PROMEDIO POR COMPONENTE	H
-7,03	0,000	0,67	1,07	0,70	1,06	12. OCUPADOS EN LA VIVIENDA	L
-8,46	0,000	1,11	1,67	0,64	1,24	15. RECEPTORES DE INGRESOS EN LA VIVIENDA	O
-8,78	0,000	385,89	699,78	288,95	663,48	7. INGRESO TOTAL DE LA FAMILIA	G

El *grupo 2* (cuadro 6) está conformado por viviendas que están habitadas por un alto número de personas (8) con relación a la media general (4-5), por lo que también resulta alto el número de receptores de ingresos (3), de ocupados (2) e inactivos (4) promedio en la vivienda; a la vez, el ingreso total promedio de la familia ($1093.20) resulta más elevado que el de la media general ($699.78) y la edad promedio de los desocupados del grupo (26) es inferior a la de la media general (35-36).

Es decir que este grupo está conformado por viviendas con un número grande de habitantes que, en conjunto, logran generar un ingreso promedio mayor al de la media general y donde los desocupados, con una edad promedio sustancialmente inferior a la media general, poseen un nivel de instrucción no muy alto (primaria completa). Las variables ilustrativas indican a su vez que también son importantes para este grupo las variables Estado Civil y Relación de Parentesco con el Jefe de la Familia, los valores de éstas indicarían que una proporción importante de los desocupados del grupo serían solteros e hijos del jefe de familia.

Cuadro 6: Caracterización del grupo 2

CLASE 2 / 3 (PESO = 35; EFECTIVO = 35) - IDENT. aa2a -							
V.TEST	PROB.	MEDIAS		DESVIO ESTANDAR		VARIABLES CARACTERISTICAS	
		CLASE	GENERAL	CLASE	GENERAL	NUM. ETIQUETA	IDEN
9,75	0,000	8,23	4,54	2,61	2,48	11. PERSONAS EN LA VIVIENDA	K
8,08	0,000	3,20	1,67	1,43	1,24	15. RECEPTORES DE INGRESOS EN LA VIVIENDA	O
7,53	0,000	2,29	1,07	1,14	1,06	12. OCUPADOS EN LA VIVIENDA	L
7,46	0,000	4,14	2,03	2,19	1,85	14. INCATIVOS	N
4,88	0,000	2,91	2,10	1,18	1,10	41. RELACIÓN	RELA
3,88	0,000	1093,20	699,78	581,33	663,48	7. INGRESO TOTAL DE LA FAMILIA	G
3,68	0,000	1,80	1,44	0,71	0,64	13. DESOCUPADOS EN LA VIVIENDA	M
2,79	0,003	2,40	2,00	0,99	0,94	16. PARA QUE BUSCA TRABAJO	SPQ
2,43	0,008	3,57	3,05	1,48	1,39	10. HABITACIONES	J
-2,47	0,007	7,49	8,71	2,72	3,25	5. AÑOS DE EDUCACION	E
-3,22	0,001	1,66	2,22	0,95	1,13	43. ESTADO CIVIL	CIVI
-4,03	0,000	26,06	35,44	11,06	15,25	9. EDAD	I

El *grupo 3* (cuadro 7) está caracterizado por tener ingresos familiares ($1506.8) y por componente ($437.60) superiores al promedio general ($699.78 y $167.10 respectivamente), a la vez que está conformado por desocupados con un nivel de educación sustancialmente superior al promedio general.

Cuadro 7: Caracterización del grupo 3

CLASE 3 / 3 (PESO = 30; EFECTIVO = 30) - IDENT. aa3a -							
V.TEST	PROB.	MEDIAS		DESVIO ESTANDAR		VARIABLES CARACTERISTICAS	
		CLASE	GENERAL	CLASE	GENERAL	NUM. ETIQUETA	IDEN
10,25	0,000	437,60	167,10	188,85	157,49	8. INGRESO PROMEDIO POR COMPONENTE	H
7,26	0,000	1506,80	699,78	876,96	663,48	7. INGRESO TOTAL DE LA FAMILIA	G
6,40	0,000	12,20	8,71	2,51	3,25	5. AÑOS DE EDUCACION	E
3,62	0,000	3,90	3,05	1,80	1,39	10. HABITACIONES	J
3,60	0,000	209,17	92,19	337,30	193,67	4. INGRESO TOTAL DEL INDIVIDUO	D
3,18	0,001	2,50	2,00	1,06	0,94	16. PARA QUE BUSCA TRABAJO	SPQ
2,38	0,009	2,17	1,67	1,10	1,24	15. RECEPTORES DE INGRESOS EN LA VIVIENDA	O
-2,90	0,002	1,13	2,03	1,06	1,85	14. INCATIVOS	N
-2,91	0,002	15,13	22,32	10,97	14,75	6. ANTIGUEDAD EN MAR DEL PLATA	F
-3,07	0,001	1,73	1,90	0,44	0,33	37. ASISTIO A LA ESCUELA	ASIS

Por lo tanto, los tres grupos tienen características interesantes que los diferencian. El tercer grupo, compuesto por el 15% de los desocupados, sería en el que los individuos sin ocupación se encontrarían en una mejor situación socioeconómica presente y con mejores perspectivas para el futuro. Esto es debido a que pertenecen a familias con ingresos relativamente altos y poseen un nivel educativo importante.

El segundo grupo, compuesto por el 19% de los desocupados, enfrentaría una situación socioeconómica relativamente buena. Esto estaría explicado por la contención que encontraría esta persona en su familia. Esta está compuesta por un elevado número de personas, de las cuales varias aportan para poder conformar, en conjunto, un ingreso familiar relativamente "alto". Por otro lado, una proporción importante de los desocupados del grupo está compuesta por solteros, hijos del jefe de familia y jóvenes.

El primer grupo, el más grande (65%), sería el que tiene una situación socioeconómica más comprometida, al pertenecer a familias que tienen un ingreso medio menor a $400, con pocas personas trabajando (ocupados) y percibiendo ingresos. A la vez, los desocupados de este grupo tienen una edad promedio alta, muchos de ellos son jefes de familia, están "casados" y con un nivel educativo "bajo".

3.4. Posicionamiento de los grupos dentro del plano factorial

La división en clases consiste en un corte más o menos arbitrario de un espacio continuo. El análisis de los ejes principales permite, además de visualizar las posiciones relativas de las clases dentro del espacio, poner en evidencia ciertas "trayectorias" marcadas por la discontinuidad de las clases. Es interesante proyectar las modalidades activas (figura 2), los centros de gravedad de los grupos (figura 3) y los individuos de cada grupo (figura 4) sobre el primer plano factorial.

El soporte visual (analógico) permite apreciar las distancias entre las clases. Por otra parte, la posición de cada individuo indicado por el número de esa clase (figura 4) permite representar la densidad y la dispersión de las clases dentro del plano factorial.

La utilización conjunta de análisis factorial y de clasificación da lugar a conocer no sólo la realidad de las clases sino también su posición relativa, forma, densidad y dispersión. De esta manera, las dos técnicas se validan mutuamente.

Si observamos la figura 4, encontramos que el grupo 1 tiene mayor densidad y sus individuos se encuentran menos dispersos que los de los grupos 2 y 3. Situación que se hace más evidente si cotejamos esta figura con la de los centros de gravedad de los grupos (figura 3).

Comparando las figuras 2 (ACP) y 4 (AC) vemos cómo las variables que definen los ejes en el ACP son las que caracterizan a los grupos en el AC. Esto permite aclarar algunas situaciones del ACP. Así, analizando el significado de las variables en términos socioeconómicos potenciales —positivo o negativo— para el individuo desocupado o su familia (cuadro 8) puede concluirse que: los efectos negativos determinaron la conformación del grupo 1, los medio–atenuante la del grupo 2 y los positivos la del grupo 3.

Cuadro 8: Significado socioeconómico de las variables en las figuras 2 y 4

Variables relacionadas con	Efecto socioeconómico	Forman parte del Grupo
Transcurso del tiempo	Negativo	1
Tamaño del grupo familiar	Medio – Atenuante	2
Nivel educativo del individuo	Positivo	3
Desocupados e inactivos en la vivienda	Negativo	1

4. Reflexiones finales

De los resultados del trabajo se pueden realizar dos tipos de reflexiones. Una referente a la utilización de métodos multivariados y la otra al análisis de los desocupados del PGP.

Con relación a la primera:

La aplicación de los métodos de ACP y AC a la información de los desocupados del Pdo. de Gral. Pueyrredon permitió ver la relevancia de estos métodos como herramienta para complementar el análisis de datos cuantitativos y cualitativos.

El análisis de componentes principales posibilitó reducir el número de variables de 15 a 8 (G, K, L, O, E, H, M y N), permitiendo fijar la atención en las características más relevantes de los individuos.

El análisis de cluster permitió, por un lado, mostrar cómo las características de los desocupados y su familia resultaban relevantes para la conformación de grupos con distintos perfiles y, por otro, dar una descripción más clara de tales características.

Con relación al análisis de los desocupados del PGP se encontró que:

Si juntamos a los individuos sin ocupación de acuerdo a sus características más relevantes, se conforman tres grupos con elementos diferenciales. Dos de estos grupos tendrían una situa-

ción socioeconómica relativamente buena, debido a que su problemática tendría contención en el seno de sus familias. Mientras que el tercer grupo, que es el más numeroso (65% de los desocupados), tendría una situación socioeconómica más comprometida, al pertenecer a familias con bajos ingresos, y pocos miembros ocupados y percibiendo ingresos. A la vez, los desocupados de este último grupo tienen una edad promedio alta, muchos de ellos son jefes de familia, están "casados" y con un nivel educativo "bajo".

El tamaño del grupo familiar al que pertenece el desocupado tiene un efecto atenuante de su problema socioeconómico; así como el nivel educativo de la familia estaría vinculado con la situación socioeconómica familiar, dado que el grupo de mayores ingresos es a su vez el de mayor nivel educativo.

Por lo tanto, a partir de este tipo de resultados queda clara la importancia de los métodos multivariados al permitir diferenciar, dentro de un conjunto de individuos, grupos con características particulares. Este aporte resulta de sustancial importancia a la hora de desarrollar una política, dado que da lugar a una clara identificación de distintos grupos objetivo hacia los que enfocar esta política.

Bibliografía

Crivisqui, E. (1997), *Presentación de Métodos de Clasificación*. Programa PRESTA, Universidad Nacional del Centro – Université Libre de Bruxelles.

Crivisqui, E. (1997), *Presentación del Análisis de Componentes Principales*. Programa PRESTA, Universidad Nacional del Centro – Université Libre de Bruxelles.

Crivisqui, E. y Villamonte, G. (1997), *Presentación de los Métodos de Análisis Factorial de Correspondencias Simples y Múltiples*. Programa PRESTA, Universidad Nacional del Centro – Université Libre de Bruxelles.

Lacabana, M. y otros (1997), *Mar del Plata en transición, Mercado de trabajo local y estrategias familiares*, Universidad Nacional de Mar del Plata.

Lebart, L., Morineau, A. y Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, París, Dunod.

Manly, B. (1988), *Statistical, a primer*, New York, Chapman and Hall Ltd.

Anexo

VARIABLES ACTIVAS

A	Tiempo que hace que el individuo está buscando trabajo (en meses) (1)
B	Tiempo transcurrido desde que dejó la última ocupación (en meses) (2)
C	Cantidad de ocupados del establecimiento de su último trabajo (3)
D	Ingreso total del individuo percibido en el mes septiembre (4)
E	Años de educación (5)
F	Tiempo que hace que vive en Mar del Plata, expresado en años (6)
G	Ingreso total de la familia (7)
H	Ingreso promedio por componente familiar (8)
I	Edad del individuo (9)
J	Cantidad de habitaciones de la vivienda (10)
K	Cantidad de personas que viven en la vivienda (11)
L	Cantidad de ocupados que habitan en la vivienda (12)
M	Cantidad de desocupados que habitan en la vivienda (13)
N	Cantidad de personas inactivas que habitan en la vivienda (14)
O	Receptores de ingreso en la vivienda (15)

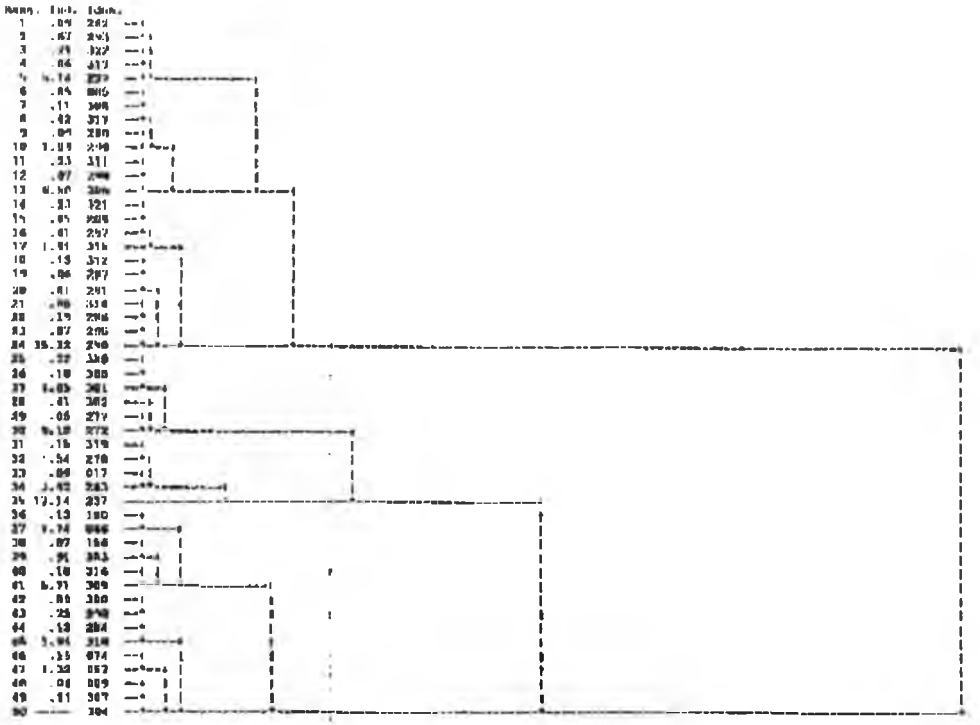
VARIABLES ILUSTRATIVAS

FALT	No encuentra trabajo porque no hay
TEMP	Temporalidad de la relación laboral
VALE	Recibe vales o tiket
LEER	Sabe leer y escribir
ESCU	Asistió a la escuela
RELA	Relación de parentesco con el jefe de familia
SEXO	Sexo
CIVI	Estado civil

Clasificación jerárquica descripción de los 49 nodos de índices más elevados.

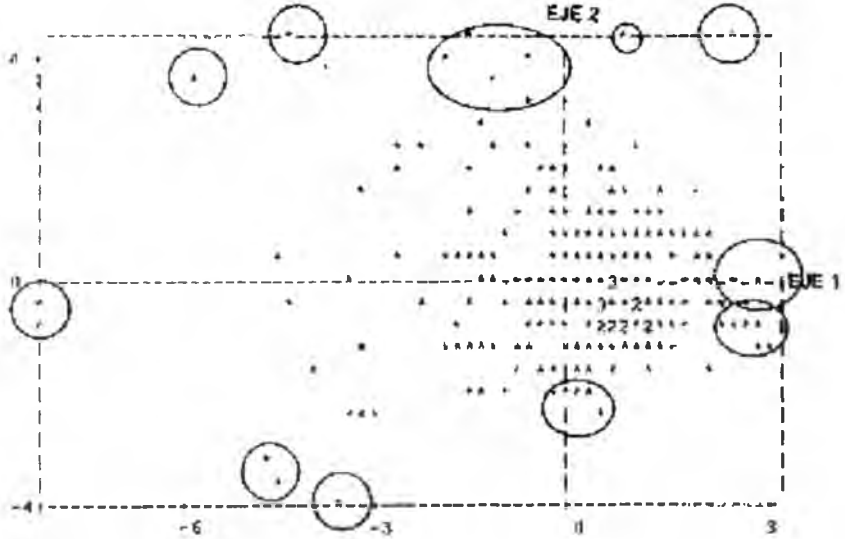
NUM	PRIM	BENJ	BHE	PESO	INDICE	HISTOGRAMA DE LOS INDICES DE NIVEL
323	272	277	7	7.00	0.00285	*
324	308	5	6	6.00	0.00294	*
325	297	289	11	11.00	0.00314	*
326	291	287	14	14.00	0.00322	*
327	307	89	3	3.00	0.00355	*
328	227	313	10	10.00	0.00374	*
329	290	295	10	10.00	0.00382	*
330	303	155	4	4.00	0.00421	*
331	306	298	10	10.00	0.00433	*
332	288	314	10	10.00	0.00442	*
333	292	300	4	4.00	0.00500	*
334	293	262	8	8.00	0.00502	*
335	263	17	3	3.00	0.00513	*
336	299	280	12	12.00	0.00543	*
337	301	305	8	8.00	0.00589	*
338	304	327	5	5.00	0.00610	*
339	317	324	14	14.00	0.00638	*
340	318	284	7	7.00	0.00667	*
341	66	180	2	2.00	0.00727	*
342	326	312	21	21.00	0.00728	*
343	278	319	5	5.00	0.00884	*
344	57	74	2	2.00	0.00885	*
345	309	316	4	4.00	0.01024	*
346	329	332	20	20.00	0.01086	*
347	328	322	13	13.00	0.01232	*
348	337	320	12	12.00	0.01269	*
349	325	321	14	14.00	0.01331	*
350	331	311	15	15.00	0.01334	*
351	340	333	11	11.00	0.01469	*
352	323	302	14	14.00	0.02371	*
353	336	339	26	26.00	0.02440	*
354	335	343	8	8.00	0.03132	**
355	315	349	19	19.00	0.03508	**
356	347	334	21	21.00	0.03859	**
357	346	342	41	41.00	0.04673	**
358	345	330	8	8.00	0.05279	***
359	352	348	26	26.00	0.06105	***
360	338	344	7	7.00	0.07659	***
361	350	353	41	41.00	0.09225	****
362	358	341	10	10.00	0.10102	****
363	357	355	60	60.00	0.11069	*****
364	360	351	18	18.00	0.11221	*****
365	237	354	10	10.00	0.20964	**********
366	361	356	62	62.00	0.29752	**************
367	364	362	28	28.00	0.33088	***************
368	363	366	122	122.00	0.37632	*****************
369	365	359	36	36.00	0.53174	************************
370	367	369	64	64.00	0.99285	**
371	370	368	186	186.00	2.04589	**
Suma de los indices de nivel = 5.85577						

DENDOGRAMA (Índices en porcentajes de la suma de los índices: 5.79282 min = .05% / max = 35.32%)



FIGURAS

FIGURA 1: POSICIÓN DE LOS INDIVIDUOS



IDENTIFICACIÓN DE LOS PUNTOS (186 INDIVIDUOS)
* : UN SOLO PUNTO
N : N PUNTOS SUPERPUESTOS X : 10 PUNTOS SUPERPUESTOS O MÁS

FIGURA 2: REPRESENTACIÓN DE LAS VARIABLES ACTIVAS E ILUSTRATIVAS DENTRO DEL PLANO DE LOS DOS PRIMEROS COMPONENTES PRINCIPALES, Z_1 Y Z_2

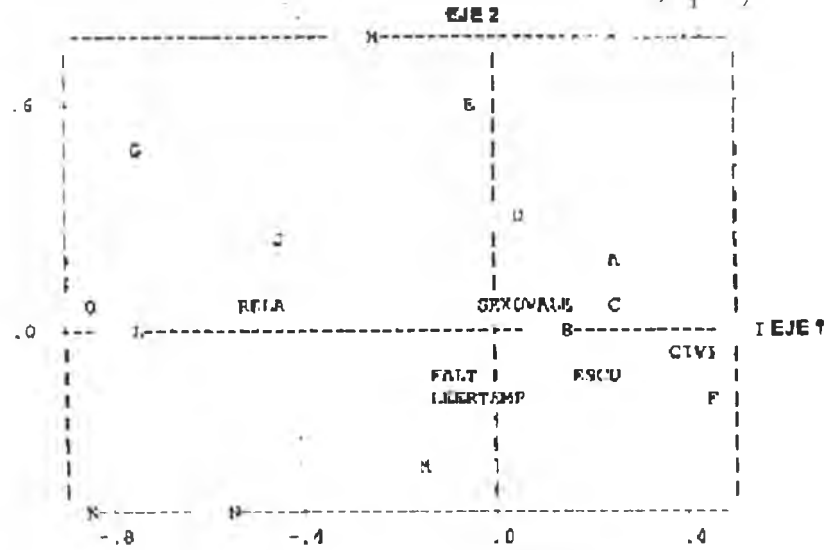


FIGURA 3: CENTROS DE GRAVEDAD DE LAS CLASES

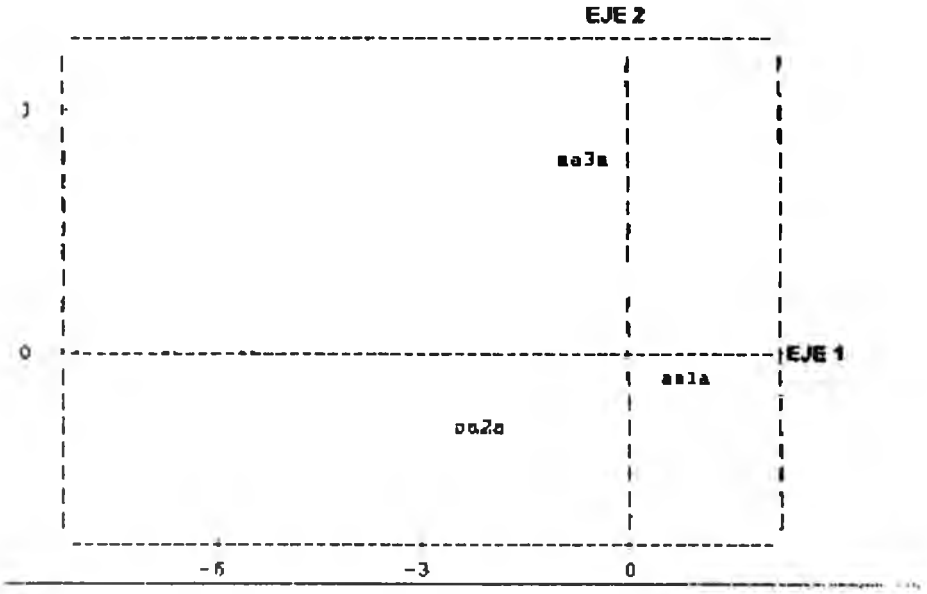
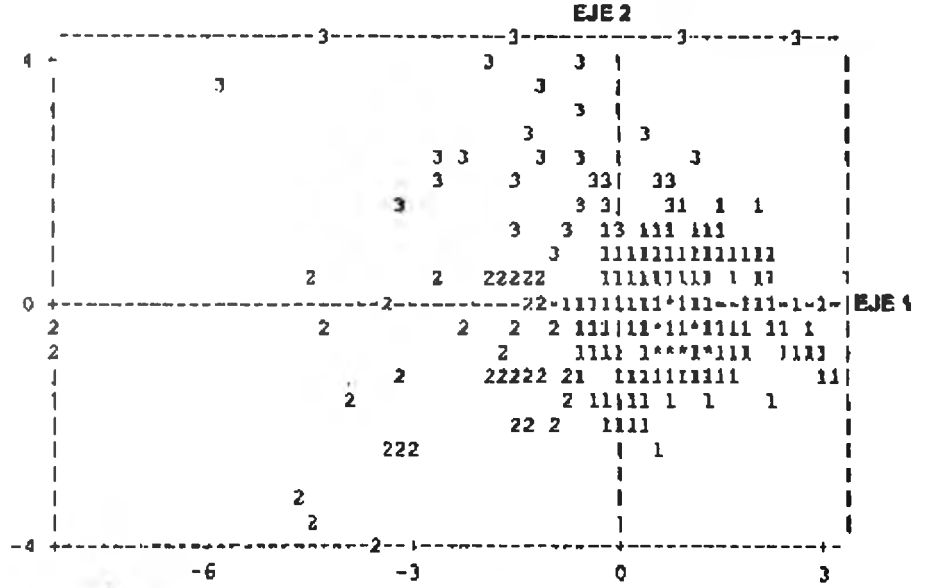


FIGURA 4: IDENTIFICACIÓN DE LOS PUNTOS (CORTE DEL ÁRBOL EN 3 CLASES)



* : VARIOS PUNTOS SUPERPUESTOS